Blue text on a black background

Description automatically generated

|  |
| --- |
| MIS 587: Business Intelligence Project Report  subtitle text here  Eller College of Management |

Flight Delay Analysis

Contents

[Chapter 1: Introduction 3](#_Toc165660342)

[1.1 Global Airlines Market 3](#_Toc165660343)

[1.2 Business Problems 3](#_Toc165660344)

[Chapter 2: Dataset Description 5](#_Toc165660345)

[2.1 Dimension Tables 6](#_Toc165660346)

[2.2 Fact Table 7](#_Toc165660347)

[Chapter 3: Data Warehouse Design and Implementation 8](#_Toc165660348)

[3.1 4-Step Dimensional Modeling Process 8](#_Toc165660349)

[3.2 Star Schema 8](#_Toc165660350)

[3.3 Warehouse Design details 9](#_Toc165660351)

[3.4 Data Warehouse Implementation in SQL Server 10](#_Toc165660352)

[3.5 ETL 12](#_Toc165660353)

[Chapter 4: Data Preparation 16](#_Toc165660354)

[4.1 Missing Values Check 16](#_Toc165660355)

[4.2 Conversion of Date Format 16](#_Toc165660356)

[4.3 Checking for Outliers 16](#_Toc165660357)

[Chapter 5: Data Exploration 17](#_Toc165660358)

[5.1 Statistical Summary 17](#_Toc165660359)

[5.2 Outlier Identification and Management 17](#_Toc165660360)

[5.3 Visual Exploration 19](#_Toc165660361)

[Chapter 6: Data Analysis and Results 21](#_Toc165660362)

[6.1 Airport Level Analysis 21](#_Toc165660363)

[6.2 Airline Level Analysis 24](#_Toc165660364)

[Chapter 7: Business Implications 26](#_Toc165660365)

[Chapter 8: References 28](#_Toc165660366)

# Chapter 1: Introduction

## 1.1 Global Airlines Market

The global airline market, valued at approximately $520 billion in 2023, is projected to nearly double, reaching around $1.1 trillion by 2030. Flight delays cost U.S. economy approximately $32.9 billion annually, with airlines and airports bearing $8.3 billion and $2 billion of those costs, respectively (FAA Report). The aim is to provide an in-depth analysis of departure and arrival delay patterns, draw insights, and suggest recommendations to airports and airlines.​

## 1.2 Business Problems

**Determining regional disparities**

* **Analysis**: Analyze historical delay data segmented by geographical regions to identify patterns and trends in delay times across different areas.
* **Action:** Develop tailored mitigation strategies for regions experiencing longer delay times, such as optimizing staffing levels, equipment deployment, and infrastructure upgrades.
* **Why:** Regional variations in delay times impact operational efficiency and passenger satisfaction, necessitating targeted interventions to address disparities.
* **Benefit:** Implementing region-specific strategies can lead to more equitable service provision, improved on-time performance, and enhanced passenger experience, ultimately bolstering the reputation of airlines and airports in those regions.

**Airport Analysis**

* **Analysis**: Identify the root causes of delays at high-impact airports through in-depth analysis of delay patterns, contributing factors, and operational inefficiencies.
* **Action**: Implement targeted interventions at identified airports, such as operational enhancements, infrastructure investments, and collaborative efforts with airlines and regulatory authorities.
* **Why**: Addressing delay hotspots at key airports is crucial for reducing system-wide disruptions and improving overall efficiency.
* **Benefit**: By prioritizing interventions at high-impact airports, stakeholders can mitigate delays, enhance service reliability, and optimize resource allocation, leading to cost savings and improved passenger satisfaction.

**Airlines Analysis**

* **Analysis**: Investigate the operational practices and performance metrics of airlines responsible for most departure delays to identify underlying causes and areas for improvement.
* **Action**: Collaborate with the identified airlines to implement measures aimed at improving operational efficiency and on-time performance, such as optimizing turnaround processes and scheduling protocols.
* **Why**: Understanding the disproportionate influence of select airlines on departure delays is essential for targeted resource allocation and collaborative interventions to enhance punctuality.
* **Benefit**: By targeting resources to improve efficiency and collaborating with airlines, stakeholders can reduce departure delays, improve service reliability, and enhance the overall passenger experience, thereby strengthening the competitiveness and reputation of the aviation sector.

**Temporal Patterns in Delay Times**

* **Analysis:** Analyze historical data to identify temporal patterns in delay times, focusing on peak periods of delays such as evenings and overnight hours.
* **Action:** Implement proactive measures to manage peak periods effectively, such as adjusting scheduling protocols, optimizing staffing levels, and deploying resources strategically.
* **Why:** Tailoring operational strategies to address peak delay periods can minimize disruptions and improve service reliability.
* **Benefit:** By managing temporal patterns of delay times, stakeholders can optimize resource utilization, reduce operational costs, and enhance passenger satisfaction by ensuring more predictable and punctual flight schedules.

**Monthly Variations**

* **Analysis:** Analyze historical data to identify temporal patterns in delay times, focusing on peak periods of delays such as evenings and overnight hours.
* **Action:** Implement proactive measures to manage peak periods effectively, such as adjusting scheduling protocols, optimizing staffing levels, and deploying resources strategically.
* **Why:** Tailoring operational strategies to address peak delay periods can minimize disruptions and improve service reliability.
* **Benefit:** By managing temporal patterns of delay times, stakeholders can optimize resource utilization, reduce operational costs, and enhance passenger satisfaction by ensuring more predictable and punctual flight schedules.

# Chapter 2: Dataset Description

Our dataset consists of 280K records for analysis. Description of all the attributes used for analysis:

* **FlightID**: Unique flight identifier.
* **FlightDate**: Scheduled departure date of the flight.
* **Origin**: Departure airport code.
* **Dest**: Arrival airport code.
* **Cancelled**: Flight cancellation indicator (1 for yes, 0 for no).
* **Diverted**: Flight diversion indicator (1 for yes, 0 for no).
* **CRSDepTime**: Scheduled departure time of the flight.
* **DepTime**: Actual departure time of flight.
* **DepDelay**: Departure delay in minutes.
* **ArrTime:** Actual arrival time.
* **AirTime**: Flight duration.
* **CRSElapsedTime**: Scheduled flight duration.
* **ActualElapsedTime**: Actual flight duration.
* **Distance**: Total flight distance.
* **Year**: Year of the flight.
* **Operating\_Airline**: Operating airline.
* **Flight\_Number\_Operating\_Airline**: Operating airline's flight number.
* **OriginAirportID**: Origin airport identifier.
* **DestAirportID**: Destination airport identifier.
* **TaxiOut**: Time from gate departure to takeoff.
* **TaxiIn**: Time from landing to gate arrival.
* **CRSArrTime**: Scheduled arrival time.
* **ArrDelay**: Arrival delay in minutes.

## 2.1 Dimension Tables

**1. Airline Dimension**

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Airline ID | Unique identifier for an airline. |
| Airline Name | Name of the airline. |

**2. Airport Dimension**

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Airport ID | Unique identifier for an airport. |
| Airport City | City where the airport is located. |
| Airport Code | IATA airport code. |
| State | State where the airport is located. |

**3. Time Dimension**

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Time ID | Unique identifier for a specific time record. |
| Hour | Hour of the day. |
| Minute | Minute of the hour. |

**4. Date Dimension**

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Flight Date | Date of the flight. |
| Quarter | Quarter of the year. |
| Month | Month of the year. |
| Day of Month | Day of the month. |
| Day of Week | Day of the week. |

## 2.2 Fact Table

**Flights Fact**

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Flight ID | Unique flight identifier. |
| Origin Airport ID | References Airport Dimension. |
| Destination Airport ID | References Airport Dimension. |
| Flight Date | References Date Dimension. |
| Airline ID | References Airline Dimension. |
| Scheduled Departure Time ID | References Time Dimension. |
| Actual Departure Time ID | References Time Dimension. |
| Scheduled Arrival Time ID | References Time Dimension. |
| Actual Arrival Time ID | References Time Dimension. |
| Departure Delay | Minutes delayed for departure. |
| Cancelled | Indicates if the flight was cancelled. |
| Diverted | Indicates if the flight was diverted. |
| Scheduled Time of Flight | Scheduled duration of the flight. |
| Actual Time of Flight | Actual duration of the flight. |
| Distance | Distance covered by the flight. |
| Arrival Delay | Minutes delayed on arrival. |
| Taxi Out | Time taken from gate departure until takeoff. |
| Taxi In | Time taken from landing to gate arrival. |

# Chapter 3: Data Warehouse Design and Implementation

## 3.1 4-Step Dimensional Modeling Process

To support our analysis of flight delays for airlines and airports, we employed a dimensional modeling process centered around the operational process of flight operations. The modeling followed a structured 4-step approach:

1. **Choose Business Process:** Our selected business process analyzes flight delays. The reason for choosing this process is to provide valuable insights into operational efficiency for airlines and airports, which can lead to improved customer satisfaction and optimized operational planning.
2. **Declare the Grain:** The grain of our data warehouse is defined at the level of individual flights. This allows us to examine the specifics of each flight, offering a detailed view of the operational aspects contributing to delays.
3. **Identify Dimensions:** The identified dimensions are:

* Airport Dimension: Includes attributes such as Airport ID, City, Code, and State, which provide geographical and operational context for the analysis.
* Airline Dimension: Contains information about the airlines, with attributes like Airline ID and Name, enabling performance comparisons across different carriers.
* Date Dimension: Comprises temporal attributes like Flight Date, Quarter, Month, Day of Month, and Day of Week, essential for trend analysis.
* Time Dimension: Includes the hour and minute components for various timestamps related to the flights, crucial for understanding the patterns in flight delays.

1. **Identify Facts:** The Flights Fact table centralizes key metrics for each flight, including departure and arrival delays, cancellation status, flight distance, and taxi times. It links to the specified dimensions to enable comprehensive analysis.

## 3.2 Star Schema

The Star Schema of our data warehouse centers on the Flights Fact Table which is linked to four Dimension Tables: Airport Dimension, Airline Dimension, Date Dimension, Time Dimension

**A screenshot of a computer

Description automatically generated**

## 3.3 Warehouse Design details

**Role Playing Dimension**

In the design of our data warehouse, we have employed role-playing dimensions, which are used to reference different business scenarios using the same dimension structure. This approach brings both efficiency and depth to our analysis. Specifically, we have two key role-playing dimensions:

* **Time Dimension**: The Time Dimension is utilized in multiple contexts within the Flights Fact table, allowing us to track different time-specific events for each flight without duplicating dimensional data. Each flight has multiple time points, such as scheduled departure and actual departure times.
* **Airport Dimension as a Role-Player**: The Airport Dimension also plays dual roles, representing both the starting point and the destination for each flight. This singular dimension contains all necessary airport data but is linked twice to the Flights Fact table - once as the Origin Airport ID and again as the Destination Airport ID.

By implementing these role-playing dimensions, our data warehouse can handle multiple time-related and location-related aspects of flights efficiently.

## 3.4 Data Warehouse Implementation in SQL Server

The data warehouse was implemented using SQL Server, reflecting the design of our Star Schema. We established tables corresponding to our schema:

* Dimension Tables: Created to hold descriptive data, these include the Airline Dimension, Airport Dimension, Date Dimension, and Time Dimension tables, each with primary keys like Airline ID, Airport ID, Flight Date, and Time ID.
* Fact Table: The Flights Fact table was set up to store transactional data like departure and arrival times, and delay durations, linked to dimension tables through foreign keys.

Screenshots from SQL Server demonstrate the tables and their relationships, confirming the successful implementation of our data warehouse structure.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

## 3.5 ETL

**Airport Dimension**

A screenshot of a computer

Description automatically generated

The ETL process for the Airport Dimension in our data warehouse consists of the following steps:

**Excel Source**: The process begins by importing a dataset with over 282,000 records from an Excel file, which contains detailed flight information.

**Multicast Transformation**: This step utilizes a Multicast transformation to split the data into multiple streams. This approach enables parallel processing and efficient data manipulation.

**Union All Transformation**: Subsequently, the Union All component merges data streams for origin and destination airports, treating each as distinct entries to capture every unique airport for analysis.

**Derived Column Transformation**:

A screenshot of a computer

Description automatically generated

Our data initially presented a concatenated format of city and state within the 'Airport City' column, such as "New York, NY." To isolate the city names for a cleaner and more focused analysis, string manipulation is implemented

Using LEFT and FINDSTRING functions in an expression to pinpoint the comma character and extract the city name portion of the string. This effectively separates the city names from the state abbreviations.

Before the transformation, the data appeared as "Phoenix, AZ" and after the application of our expression, the output was simplified to "Phoenix." This refinement ensures uniformity and precision within the 'Airport City' column, facilitating subsequent data analysis where city-specific information is required.

**Sort Transformation**: The final step in the ETL process uses a Sort transformation to eliminate duplicate records, thereby refining the list down to 227 unique airport records, ready for analysis.

**Control Flow**

A screenshot of a computer

Description automatically generated

In our data warehouse's ETL process, we've established a sequence of operations to extract, transform, and load data from various sources into our system. Each task in this process represents a specific step aimed at structuring and incorporating data into our data warehouse.

* Airport Dimension: Processes and formats airport details.
* Time and Date Dimensions: Structures temporal data for analysis.
* Airline Dimension: Prepares airline data for the related dimension table.
* Flights Fact: Consolidates key flight metrics into the fact table.

The presence of green checkmarks indicates the successful completion of each task in our ETL process. These checkmarks signify that the data extraction, transformation, and loading procedures have been executed without errors.

After the ETL processes were executed, the Fact and Dimension table’s final structure was established within the SQL Server database, as demonstrated in the query output screenshots

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# Chapter 4: Data Preparation

In the preliminary stages of data preparation, the focus was on establishing a foundation of data integrity and usability for subsequent ETL processes and analysis.

## 4.1 Missing Values Check

The dataset underwent a thorough examination for missing values across all columns. Utilizing Python's pandas library, the following code was executed to enumerate null entries:

A close up of a text

Description automatically generated

The output confirmed the absence of missing values, indicating that the dataset was complete and that imputation strategies were not required

## 4.2 Conversion of Date Format

The 'FlightDate' field, vital for any time series analysis, was originally in an object data type. For more efficient manipulation and analysis, we converted 'FlightDate' to the datetime data type using pd.to\_datetime(flightsdata['FlightDate']). This conversion enables accurate sorting, grouping, and interval calculations.

A screen shot of a computer code

Description automatically generated

## 4.3 Checking for Outliers

After the initial analysis, the identification of outliers was carried out using the Interquartile Range (IQR) method. Entries exceeding these defined limits were classified as outliers. The exclusion of such data points refined the overall dataset to 259,073 entries.

A screenshot of a computer

Description automatically generated

# Chapter 5: Data Exploration

## 5.1 Statistical Summary

The exploratory data analysis began with generating a statistical summary of the dataset to understand the central tendencies and variability of departure delays:

A close-up of a computer screen

Description automatically generated

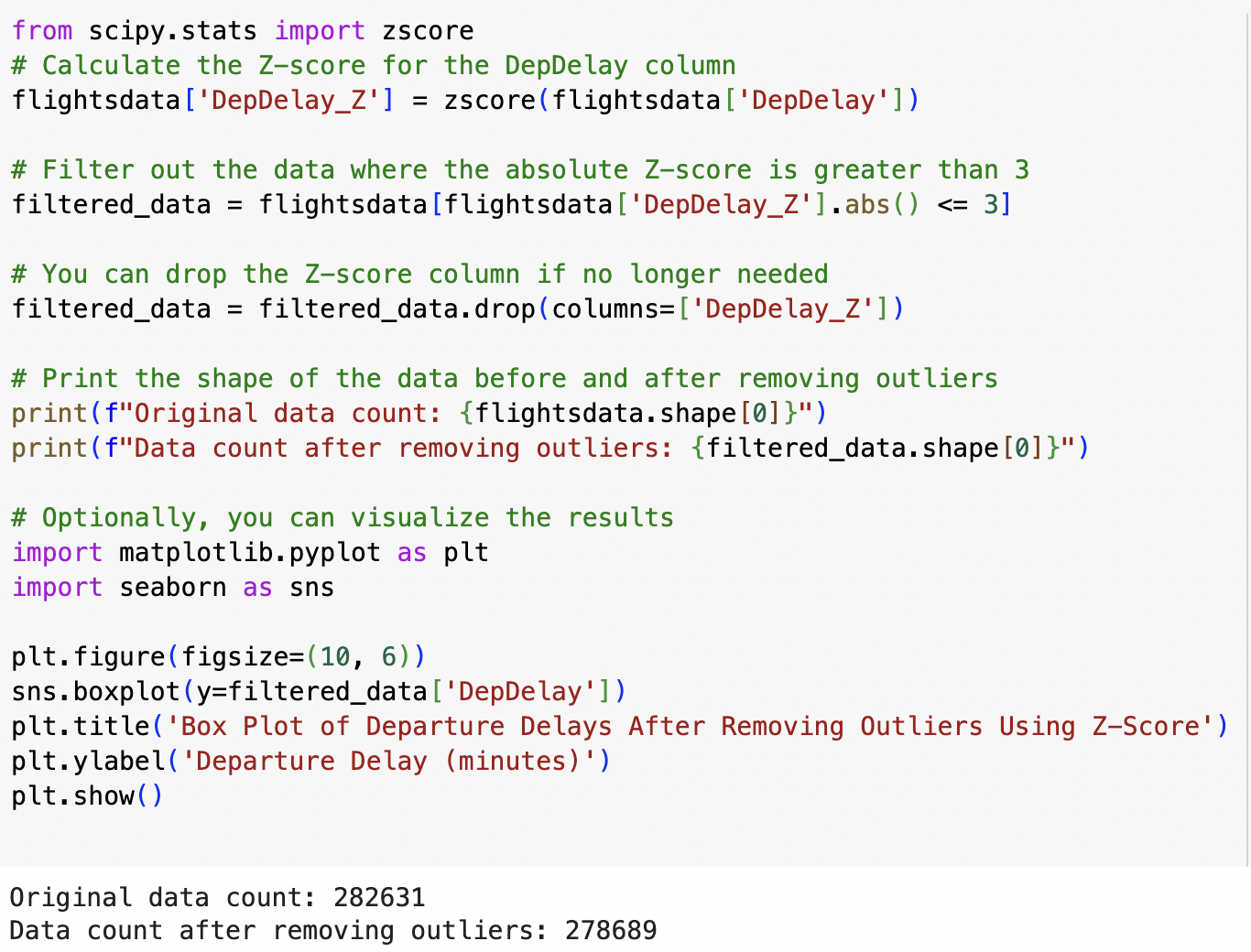
A number of numbers and a percentage

Description automatically generated with medium confidence

The descriptive statistics for DepDelay indicated an average delay of around 52 minutes. However, the range of delays was substantial, with the minimum being 1 minute and the maximum reaching up to 2545 minutes, suggesting the presence of extreme values or outliers.

## **5.2 Outlier Identification and Management**

The application of z-score calculations was undertaken to detect and address outliers within the departure delay data (DepDelay). A z-score quantifies how many standard deviations an element is from the mean, providing a basis for identifying data points that deviate significantly from the expected range. The calculation and filtration process is encapsulated in the following code:



This procedure effectively refined the dataset, reducing the data count from 282,631 to 278,689 observations.

After this cleansing, the filtered dataset's departure delays were visualized using a box plot, offering a clear representation of the central tendency and variability without the distortion of outliers.

A graph with a line and a bar

Description automatically generated with medium confidence

## **5.3 Visual Exploration**

**Analysis of Average Departure Delays by Airline**

An examination of departure delays by airline was conducted using Python’s pandas and seaborn libraries. A bar plot was created to compare the average delays:

A screenshot of a computer code

Description automatically generated

The resulting visualization highlights significant differences in average delays among airlines.

A graph with a bar chart

Description automatically generated with medium confidence

**Relationship between Departure Delay and Distance**

A scatter plot was generated to investigate the relationship between flight departure delays and the distance traveled.

A screen shot of a computer

Description automatically generated

The scatter plot displayed a dispersed set of points, indicating no clear linear relationship between the length of the delay and the distance of the flight. This suggests that other factors may have a more significant impact on the length of delays than the distance alone.

A graph of a scatter plot

Description automatically generated

# Chapter 6: Data Analysis and Results

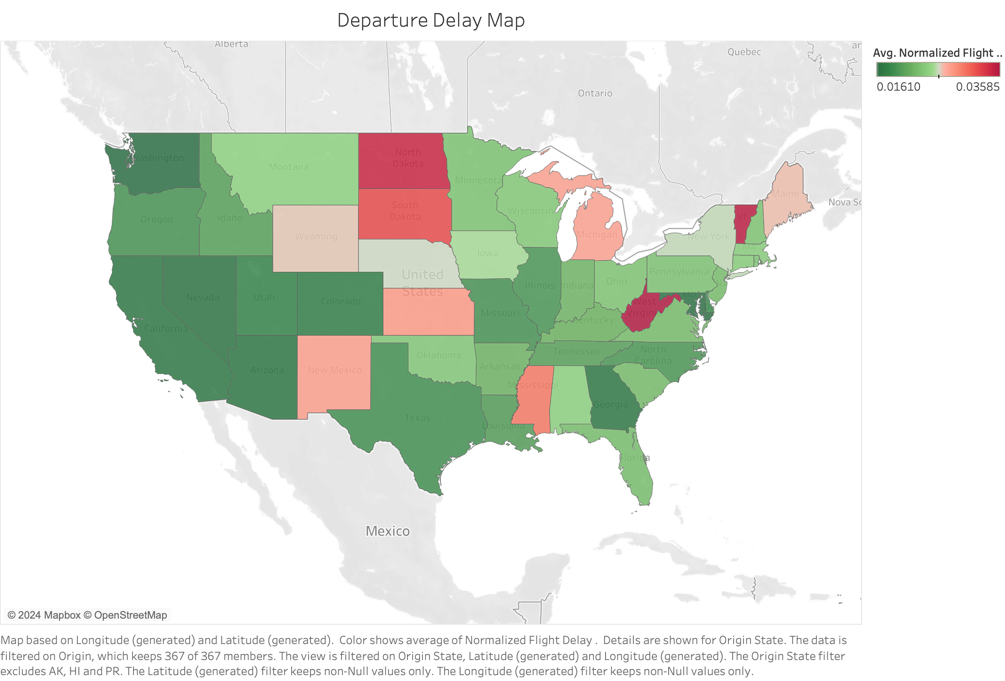
Our airline analysis focused on flight delays at the airport level and the airline level. Our methodology employed visual analytics to perform a drill down analysis on the 2022 United States airline delay data set. The types of visuals developed are a map, pareto chart, line graph, bar graph, tables, and heat map. From these visuals we were able to gain insights that impact consumer decisions, airport/airline staffing, and airline/airport operations.

## 6.1 Airport Level Analysis

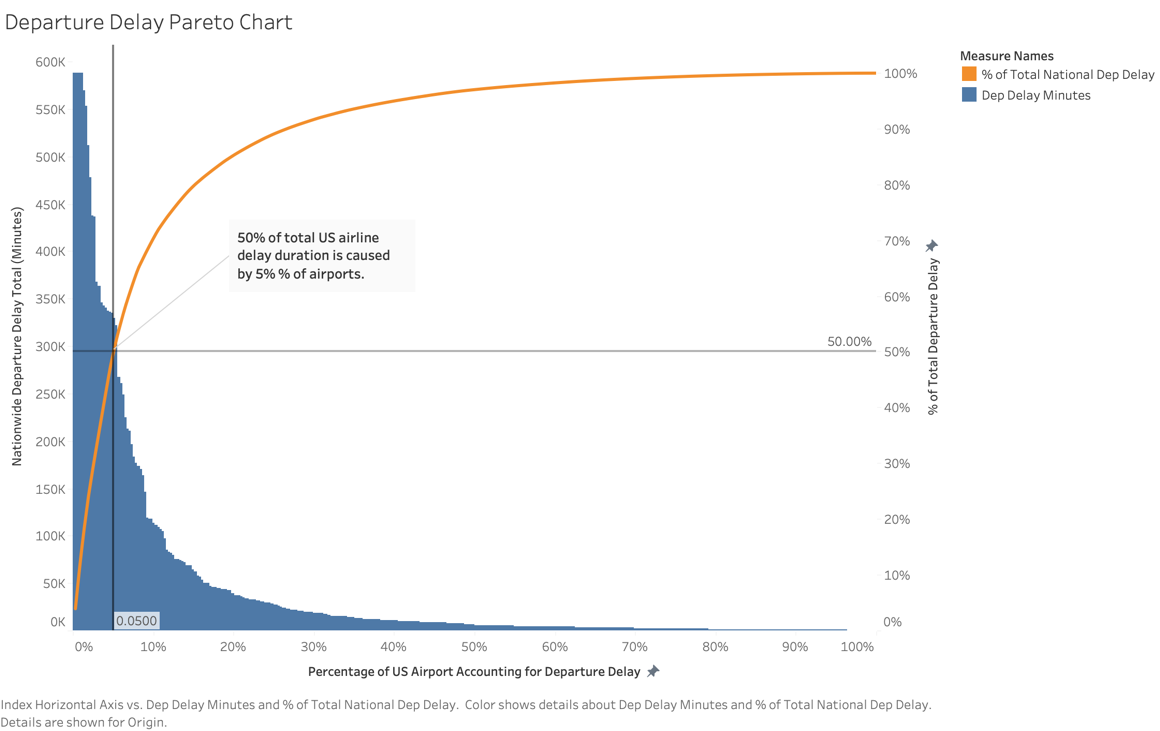
The first visual created was a map of the United States. The map contained two different data points which are the average delay time and the count of flight delays. Both were aggregated at the state level. The color of each state represents the average departure delay time with red being above a higher state average and green being a lower state average. The dots represent the count of departure delays instances that occurred in the state. The size of the dot indicates the count with larger dots corresponding to higher delay counts and vice versa.



From the visual we determined that states on the West Coast have the shortest delay times whereas states in the Midwest and East Coast have longer average delay times. Surprisingly, the states that had the highest delay counts, tended to have lower average delay times. One reason could be the small sample size of states with high averages. To combat this, we incorporated feedback from class and normalized the average departure delay data. The normalization method selected is min-max normalization. However, after completing this, there were no changes to the results in the map, further supporting our earlier conclusions.



To further investigate the state trends identified, we created a Pareto chart of all airports in the United States. The Pareto chart shown below includes a bar graph and a line graph showing two different things. The line graph represents the cumulative total of all flight delay minutes in the United States in 2022 while the bar graph shows the percentage of airports accounting for the percentile of delay minutes. From this visual, we determined that a relatively small number of airports are accounting for most delays. Specifically, 5% of airports account for 50% of total flight delays in minutes.



We then created a table to outline the 5% of airports causing 50% of the delays. From the below table we can see that 50% of delayed minutes is from only 19 airports, out of the 367 examined. In class, a piece of feedback given was to normalize or convert the delay durations into ratios. It was decided not to convert due to the values required for converting to ratios not being available in the dataset. Introducing values from outside resource posed the risk of introducing skewness and bias to the dataset which would decrease the reliability of the analysis.

A screenshot of a screen

Description automatically generated

Based on these findings, we did a further analysis into the performance of these 19 airports. We created a heatmap examining the relationship between average delay duration and time of day at each airport, which can be seen below. Across the airports, their worst delay times are on evening and overnight flights as indicated by the dark red.



We then created a line graph to investigate whether this pattern occurred across all day of a month. For this we created a line chart depicting avg departure delay vs day of the month by part of the day for the 19 airports. Consistently the blue line for evening flights is the highest or second highest line, further supporting this conclusion. A surprising find that can impact consumers is that flight delay times are lower in the middle of the month. The line graphs themselves appear to have a slight parabola shape with the longest delays occurring at the beginning and end of a month.

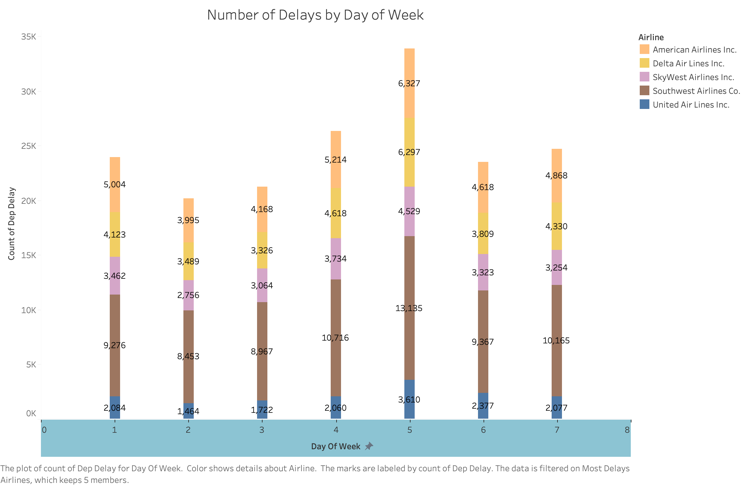
## 6.2 Airline Level Analysis

Given that each airline operates at airports across the country, our next stage of analysis focused on individual airlines. We started by creating the same pareto chart to determine the airlines responsible for the greatest proportion of flight delay minutes. The table can be seen in the figure below.

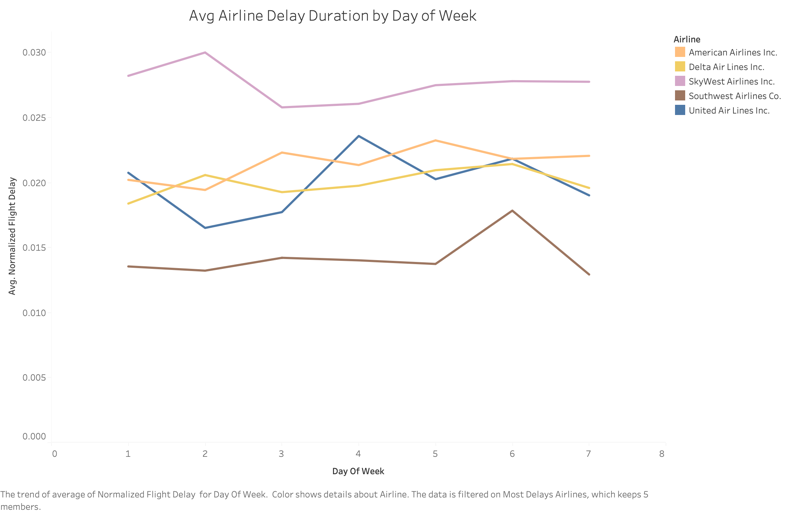
A screenshot of a table with numbers and a number of airliners

Description automatically generated

From this we can see that greater than 60% of the departure delays are caused by only 5 airlines. To further dive into this, we created a bar graph showing delay count vs day of the week. From this figure it can be seen that Southwest Airlines accounts for the largest number of delays, with their worst performing day falling on Fridays and second worst day on Thursdays.



Further analysis on these airlines was performed through a line graph of average normalized flight delay vs day of week. For this analysis delay duration was normalized using a min-max normalization technique. Surprisingly, Southwest Airlines shown by the brown line in the below figure, had the lowest average flight delay. This indicates that they have effective mitigation strategies in place to combat the large number of delays they experience. Further it poses the potential for other airlines to attempt to collaborate to improve their own delay durations.



# Chapter 7: Business Implications

In the aviation industry's dynamic landscape, data-driven insights are paramount for optimizing operational efficiency and enhancing passenger experiences. This section delves into key findings and their business implications, offering actionable strategies to drive innovation, collaboration, and sustainable growth across airports and airlines in the USA.

1. **Regional Disparities in Delay Times**
   * **Findings:** West Coast states experience shorter delay times compared to Midwest and East Coast states.
   * **Implications:** Understanding regional variations is crucial for targeted resource allocation and mitigation strategies.
   * **Application:** The system enables regional analysis of delay patterns, facilitating the allocation of resources such as staff, equipment, and infrastructure upgrades based on specific regional needs. Decision support tools can recommend tailored strategies for each region to minimize delay times and enhance operational efficiency.
2. **Concentration of Delayed Airports**

* **Findings:** A small percentage of airports account for a significant portion of total flight delays.
  + 5% of airports account for 50% of total flight delays in minutes
  + 50% of delay minutes is from only 19 airports, out of the 367 examined
* **Implications**: Prioritizing interventions at key airports can yield substantial overall improvements.
* **Application**: The system identifies high-impact airports and provides detailed insights into delay causes and patterns at these facilities. Decision support modules offer targeted recommendations for operational enhancements, resource allocation, and collaboration with airlines and regulatory authorities to address delay hotspots effectively.

1. **Temporal Patterns in Delay Times**

* **Findings**: Evening and overnight flights consistently experience the worst delay times across airports.
* **Implications**: Tailoring operational strategies to address peak delay periods can mitigate disruptions and improve service reliability.
* **Application**: Leveraging historical data on temporal delay patterns, the system assists airports in implementing proactive measures to manage peak periods effectively. Decision support functionalities enable real-time monitoring and adaptive response to anticipate and mitigate delays during critical time windows, optimizing resource utilization and passenger experience.

1. **Monthly Variation in Delay Times**

* **Findings**: Flight delay times are lower in the middle of the month compared to other periods.
* **Implications**: Identifying temporal trends allows for targeted resource allocation and scheduling adjustments.
* **Application**: Utilizing predictive analytics, the system forecasts monthly delay trends and recommends optimized resource allocation strategies accordingly. Decision support tools facilitate dynamic scheduling adjustments, staffing levels, and operational protocols to align with anticipated variations in delay times throughout the month, maximizing efficiency and minimizing costs.

1. **Departure Delays Dominated by Select Airlines** 
   * **Findings:** Departure delays in the dataset are significantly influenced by a minority of airlines. Specifically, over 60% of departure delays can be attributed to just five airlines.
   * **Implications:** This data reveals that just five airlines are responsible for over 60% of departure delays, highlighting a significant concentration of punctuality challenges within the aviation sector.
   * **Application:** Understanding this, airlines can target resources to improve efficiency, while airports and regulators can collaborate with these carriers to implement measures that enhance on-time performance, ultimately benefiting passengers and the industry.

# Chapter 8: References

* 1. Buomsoo Kim, Hong Cui , Yong Ge & Faiz Currim Data Exploration and Preprocessing
  2. SQL Server Integration Services - SQL Server Integration Services (SSIS). Learn.microsoft.com, learn.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-ver16
  3. SQL Server Technical Documentation - SQL Server. Learn.microsoft.com, learn.microsoft.com/en-us/sql/sql-server/?view=sql-server-ver16
  4. Pandas Documentation - Pandas.pydata.org, 2024, pandas.pydata.org/docs/
  5. Numpy Documentation - Numpy.org, numpy.org/doc/
  6. Power BI Documentation - Power BI. Learn.microsoft.com, learn.microsoft.com/en-us/power-bi/